Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Multi-block tensor regression for quality prediction and root cause analysis in the production of active pharmaceutical ingredients

Carlos André Munoz Lopez * **Maité Lenaerts** **
Kristin Peeters ** **Jan Van Impe** *

* *KU Leuven, Chemical Engineering Department, BioTeC+&OPTEC, Gebroeders De Smetstraat, 9000 Ghent, Belgium (e-mail: jan.vanimpe@kuleuven.be).*
** *Janssen Pharmaceutica, Technical Operations, Geel chemical production site, Janssen-Pharmaceuticalaan 3, 2440 Geel, Belgium (e-mail: kpeeter1@its.jnj.com).*

Abstract: The large scale production of active pharmaceutical ingredients (APIs) is traditionally accomplished via batch processes. Nevertheless, their inherent complexity has limited the development and application of models for the processes as well as the use of advanced online monitoring, control and optimization strategies for their continuous improvement. The quality by design (QbD) strategy, defined by regulatory agencies, has brought a practical view into this. QbD seeks to determine methods to define the critical quality attributes (CQAs) of the product in terms of the critical material attributes (CMAs) and the critical process parameters (CPPs) of the input space along the whole process. This means not only on individual batch units but also throughout the multiple steps and stages of the production process. In this contribution, data-driven modelling methods were exploited to model two synthesis steps in the large scale production of an Active Pharmaceutical Ingredient (API). First, tensor factorization was applied to train correlation based models to capture the main directions of variability for each of the studied steps. Secondly, a partial least squares (PLS) model was trained to regress the concentration of an impurity on the product onto the latent variables of the primary monitoring models as well as the CMAs of the input materials. The proposed modelling approach was applied to data from the large scale production on an API. This resulted in an accurate model, which was validated on an independent data set and which captures meaningful correlations that helped on the root cause identification for variations encountered in the CQA of the product. These results are in line with the observations made on the process operation.

Keywords: Data-driven modelling, tensors, multi-block regression, Active pharmaceutical ingredient, Quality by Design.

1. INTRODUCTION

The APIs are produced trough the sequence of several synthesis steps which in turn consist of a sequence of several unit operations. These processes are carried out to transform the raw materials into the desired products, and to separate and purify the final API. Currently, the pharmaceutical industry relies mostly on batch operations for the large scale production of the APIs. The batch operation allows for flexibility in the use of standard pieces of equipment for a diverse portfolio of products and also allows the rigorous quality control of the products. However, the complexity of the batch processes combined with the pressure to shorten the time to market has limited the progress made for modelling and control of these operations. In recent years a significant increase in the efforts have been put into data-driven modelling approaches. This thanks to the large data sets that are collected routinely and the advances in data storage and computational capacity. These models aim to capture correlations existing in the systems and use them for process monitoring control and optimization. Additionally, the introduction of the QbD strategies and the Process Analytical Technology (PAT) tools have boosted the use of models to support the pharmaceutical development and manufacturing activities (Tomba et al., 2013). However, given the high complexity of the large scale pharmaceutical production systems, developing data-driven modelling approaches that can account for the data structure and the different sources of variability requires further research and development.

Several contributions can be found in the topic of multi-block data-driven modelling. Most of the work has been focused on extending the traditional linear and bi-linear approaches, e.g., Principal Component Analysis (PCA) and Partial Least Squares (PLS). In case of batch processes, this work has been extended based on multi-way methods that require to unfold the data (Nomikos and

MacGregor, 1994; Kourti et al., 1995). Hong et al. (2014) compare the consensus PCA and the multi-block PCA (MBPCA) and proposed an improved method for progressive MBPCA. Jiang and Yan (2014) extended the traditional MBPCA approach for automated block identification based on the mutual information between variables in a plant-wide process. Jiang and Yan (2015) proposed a kernel-MBPCA method to account for non-linearities. In the case of multi-block PLS (MBPLS), most of the contributions have been on applications of the standard approach proposed by MacGregor et al. (1994). Westerhuis and Coenegracht (1997) applied this method to the wet granulation and tabletting of pharmaceutical products. Most recent methods intend to separate common from distinct variation in the blocks, e.g., OnPLS, SOPLS, and DISCO. Skotare et al. (2018) and Måge et al. (2018) discuss and compare some of these methods. However, the correct identification of these models can be troublesome due to the *a-priori* assumptions on the structure of the variance and these still require unfolding the batch data. Another challenge in multi-block regression is the selection of the variables or features from the input space to be included in the regression. In this line, Biancolillo et al. (2016) compare some standard approaches and Biancolillo et al. (2019) present a novel strategy based on covariance selection.

In contrast to most of the previous work, in this contribution, the inherent structure of the data is acknowledged and the multi-block strategy is based on the tensor factorization. The use of tensor decomposition has been presented as a more suited approach for data-driven modelling of batch processes since the original structure of the data is kept in the model (Hu and Yuan, 2009; Gurden et al., 2001). Despite the proven advantages of the tensor factorization, the number of applications for batch process modelling is relatively limited. An application of tensor-based regression methods is provided by Luo et al. (2016) who use of High Order PLS for the quality prediction on a simulated fed-batch reactor. Additionally, most applications based on tensor decomposition are focus on modelling a single blocks of data Fanaee and Gama (2016). The methods based on coupled tensor-matrix factorization and data fusion explore the multi-block characteristics of the data including tensors (Ermi et al., 2013).

In this contribution, a novel strategy is presented for multi-block tensor regression and evidence of its better performance is provided based on a real data set from the large scale production of an API. A hierarchical modelling approach is formulated. First, the individual blocks are factorized using Tucker3 and then an overall regression model is trained to predict the desired output. This strategy works based on the method presented by Munoz et al. (2018) for simultaneous data scaling and training of the tensor models. This approach guarantees that the tensor models capture and factorize the deterministic variability of the processes, which is required for the regression model. The results obtained not only demonstrate the better performance of this strategy to predict the output, but also the interpretability offered by the constructed hierarchical model. This application resulted in a model that captures meaningful input-output correlations which contributed to build a better process understanding and helped in the root cause analysis of the undesired variability. This application resulted in the formulation of control actions in order to limit the variability on the quality of the product and reduce the risk of having out of specification product.

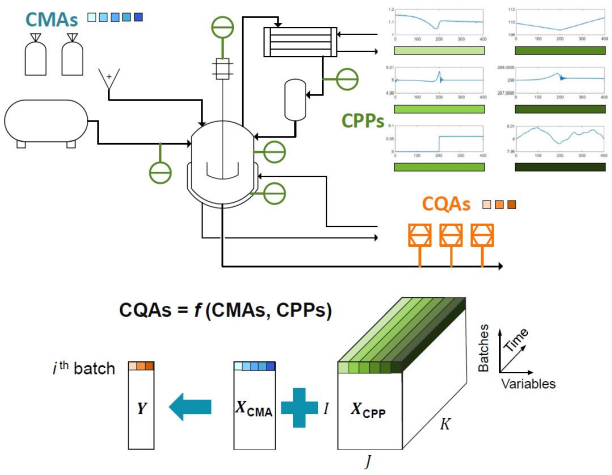

Fig. 1. Structure of the data from batch processes.

2. BATCH DATA SET AND TENSOR BASED DECOMPOSITION

Continuous data from batch processes comes naturally as a third-order tensor. This structure is shown in Fig. 1, where the dataset ($\boldsymbol{X}_{\mathbf{CPP}[I\times J\times K]}$) is a third-order tensor that has K time points measured for J variables per batch. This data can be seen as I horizontal slices of data, one per batch, which are stacked in the vertical direction. The unfolded versions of this data set are the batch-wise ($X_{(1)[I\times JK]}$), variable-wise ($X_{(2)[J\times IK]}$) and the time-wise ($X_{(3)[K\times JI]}$) matrices. In Fig. 1 the QbD approach is demonstrated on the different data sets involved in the production of an API. QbD establishes that the control strategy for the operation of the process is defined in terms of the CQAs which are a function of the CMAs and CPPs. Thus, in terms of a regression model the output space (Y) is regressed into the input space which consists of the variables contained in X_{CMA} and $\boldsymbol{X}_{\mathbf{CPP}}$. However, this is a simplified representation of the real input space. The real case involves several blocks of data, since multiple syntheses steps and multiple sub-steps are performed to complete the production of the desired API product.

The decomposition of individual tensors can be performed via multilinear methods to keep the original structure of the data. The Kronecker product ($\otimes$), given according to equation (1), is introduced to formulate the tensor factorization with a matrix notation. Tucker3 factorization is used because it offers a more flexible decomposition with an independent rank for every mode of the tensor. This is not the case with the Canonical polyadic decomposition (CPD)(Fanaee and Gama, 2016). Although CPD offers several more favorable properties such as uniqueness, preliminary studies showed that higher accuracy is achieved using Tucker3 due to its less rigid structure. Applying Tucker3 to a third-order tensor results in three factor matrices and a core tensor. This structure corresponds with the multilinear singular value decomposition (SDV) of the tensor. If the three factor matrices are orthonormal

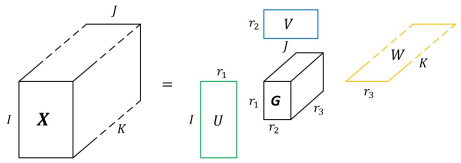

Fig. 2. Tucker3 decomposition of a third order tensor.

and are chosen as the right singular vectors of the unfolded matrices in the three directions (i.e., $X_{(1)}, X_{(2)}, X_{(3)}$) the decomposition results in a core tensor for which the slices are equivalent to the singular values (Sidiropoulos et al., 2017). In Fig. 2 Tucker3 decomposition is depicted and its formulation is presented in equation (2). The reader is referred to Sidiropoulos et al. (2017) for further details on the properties and derivations of these methods, as well as the procedures applied to compute them.

$$A_{[I_1 \times I_2]} \otimes B_{[I_3 \times I_4]} = \begin{bmatrix} a_{11}B & a_{12}B & \cdots \\ a_{21}B & a_{22}B & \cdots \\ \vdots & \vdots & \end{bmatrix}_{[I_1 I_3 \times I_2 I_4]} \tag{1}$$

$$\begin{aligned} \boldsymbol{X}(i,j,k) &\approx \sum_{l=1}^{r_1}\sum_{m=1}^{r_2}\sum_{n=1}^{r_3} \boldsymbol{G}(l,m,n)\, U(i,l)\, V(j,m)\, W(k,n) \\ X_{(1)} &= U\, G_{(1)} (W \otimes V)^T + E_{(1)} \end{aligned} \tag{2}$$

3. MULTI-BLOCK TENSOR REGRESSION

Chemical and biochemical processes normally consist of a sequence of several operations involving (bio)chemical and physical phenomena. The independent nature of these phenomena calls for independent models, but the sequential nature of the transformations impose relations that move throughout the system, transferring effects from one operation to the followings. From the data-driven modelling perspective this structure corresponds with a multi-block regression. In this contribution, the proposed strategy to perform multi-block tensor regression consists on a hierarchical modelling approach where the scores of individual Tucker3 models, i.e., one for every step involved in the process, are used as the input space for the regression of the output variable(s). This strategy is depicted in Fig. 3, where two blocks of process data from two consecutive synthesis steps, each consisting of three sub-steps, are factorized via Tucker3. The obtained 6 matrices of scores are then used together with the time-invariant data in the regression of the output space. Backward elimination is applied as part of the formulated strategy to remove those features that do not contribute significantly to the prediction of the output variable. Feature elimination improves the accuracy of the model and contributes to the identification of meaningful correlations. The selection of the features to be eliminated was done following a wrapper approach on the PLS training. Thus, at every iteration individual features were selected to be eliminated based on the improvement of the model accuracy in a leaving one out cross-validation scheme.

Tucker3 factorization of the data is applied using the simultaneous training and scaling approach proposed by Munoz et al. (2018). This method improves the ability of the Tucker3 decomposition to capture and factorize the deterministic variability of the individual steps in the batch processes. In turn, this method reduces the noise

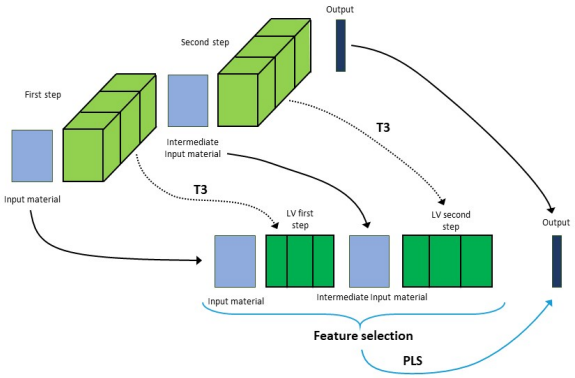

Fig. 3. Multi-block regression based on tensor factorization of batch process data.

propagation, improves the rank approximation and allows a more meaningful interpretation of the batch to batch variability captured on the scores of the decomposition. This scaling and training method does not require the data to be normalized, only mean centred. A scaling parameter P_j is applied to every variable in the data set as part of the training algorithm. The vector $P_{[1\times J]}$ is estimated to guarantee that the resulting Tucker3 model approximates evenly the variance present in every variable of the system. Equation (3) is the optimization formulation for this method, its application was completed based on the algorithmic approach presented by Munoz et al. (2018). Linear regression based on PLS is applied since the model is intended to capture the system variability around the mean trajectory which is assumed to be linear, equivalently the variance in the output is limited to a local region around the nominal value.

$$\min_{P,U,V,W} \|(X_{(2)}^T \odot P)^T - V\, G_{(2)} (W \otimes U)^T\|_F^2 + \lambda \| \frac{diag(E_{(2)} E_{(2)}^T)}{IK} - 1 \|_F^2 \tag{3}$$

4. CASE STUDY: LARGE SCALE PRODUCTION OF AN API

Historical data from two sequential synthesis steps in the production of an API were considered for this case study. Each of these steps consists of more than one sub-step, and in both cases, reaction and crystallization processes are involved. The intermediate product is isolated and the water content is tested before the intermediate product is used in the second step. The product from the second step is isolated and the quality is tested to check compliance with the quality specifications. Thus, the aim of modelling these two synthesis steps is to predict one of the quality attributes of the final product based on the process conditions and the properties of the input and intermediate materials. The multi-block model is intended also to perform root cause analysis to determine the conditions that contribute the most to the variance in the quality attribute of the product and that resulted in extreme values.

A total of 5 data sets are involved in this case study. Three matrices, two correspond to the properties of the input materials for the first and second synthesis steps and the third is the quality of the end product. The process data coming from the two synthesis steps take

the form of two overall tensors, which each, in turn, consists of multiple sub-sets, one per each phase of the synthesis (e.g., reaction, separation and crystallization). The application of the proposed multi-block approach to this data set results on factorizing the tensor data into the scores which capture the variance of the system in the batch direction. The input space for regression consists of the two input matrices as well as the two overall matrices of scores obtained from the tensor factorization. Table 1 summarizes the structure of the industrial data set used in this application. A total of 24 historical batches were used to train the model while 9 independent batches were used for validation. The validation batches were chosen to be more than 25% of the available data and these correspond with the most recent production campaign. These batches have a similar distribution on the quality attribute to that of the training batches. This allowed validating the model throughout the full range of the output variable. In Table 1 some of the variables contained in every data set are listed to offer and indication on the type of data included in the model. The dimensions of the data sets in the table are reported as the number of batches × number of variables × number of time points. The continuous data were aligned to have the same number of time points. The alignment was performed by linear compression/expansion of the data to have the same duration for every phase of the processes. The original continuous data was sampled at a frequency of 1 [1/m].

Table 1. Data set from the large scale production of an API

Data	Data structure	Dimension	Variables
Input properties	Matrix	33×10	Specific gravity, pH, Assay
First step, reaction	Tensor	$33 \times 6 \times 450$	Reactor Temp, stirring speed, Jacket temperature
First step, crystallization	Tensor	$33 \times 6 \times 300$	Reactor Temp, stirring speed, Jacket temperature
Inter. properties	Matrix	33×9	Water content, refractive index, impurities
Second step, reaction	Tensor	$33 \times 6 \times 350$	Reactor Temp, stirring speed, Vapor temperature
Second step, Separation	Tensor	$33 \times 6 \times 200$	Pressure, stirring speed, Vapor temperature,
Second step, crystallization	Tensor	$33 \times 6 \times 300$	Reactor Temp, stirring speed, Jacket temperature
Output quality	Matrix	33×1	Impurity content

5. RESULTS

5.1 Tensor factorization

Table 2 shows the results on the tensor factorization for every set of process data. The results are presented in terms of the best rank approximation found for the Tucker3 decomposition using the method for simultaneous training and scaling. The relative error on the approximation reported as well. The rank was estimated based on the leaving one out cross validation on the training data set. The approach proposed by Tomba et al. (2013) was used to reduce the search space for the best rank approximation. This approach eliminates those combinations of the trilinear rank that result in redundant solutions. From the perspective of process monitoring the Tucker3 factorization of the process data results in a set of data-driven models for each step of the process that can be adapted to operate for online multivariate monitoring of the process.

Table 2. Results on the Tucker3 factorization of each block of process data

Data	Best rank approx.	Relative error
First step, reaction	[3,4,5]	0.48
First step, crystallization	[3,3,3]	0.26
Second step, reaction	[6,5,3]	0.20
Second step, Separation	[3,2,3]	0.20
Second step, crystallization	[6,6,2]	0.37

5.2 Quality regression

Once the process data in the tensor form is factorized, the PLS regression model was trained based on the scores obtained for every part of the process and the properties of the input materials. Feature selection was applied using the wrapper algorithm based on backward elimination. The results obtained from the proposed approach were compared to results from other models, i.e., standard regression methods based on unfolded data as well as other multi-block strategies. 6 additional regression models were evaluated. First, a standard PLS model was trained based only on data from the first synthesis step. This model responded to the hypothesis that the variance on the final quality attribute of the product originated from a precursor that is formed during the first synthesis step. This model considered the CMA's of the input and the scores obtained from the tensor factorization of the first synthesis step. Secondly, three PLS models were trained based on the adverse idea that suggested the variance in the output being originated during the second synthesis step. Thus, the unfolded data from the second synthesis step was used to train two independent models of two sub-steps and one model for the overall second synthesis step. Finally, 2 extra models were trained considering a multi-block strategy to account for variance in both synthesis steps. First, a standard MBPLS (Westerhuis and Coenegracht (1997)) was trained based on the unfolded data from the second synthesis step, as well as, the scores

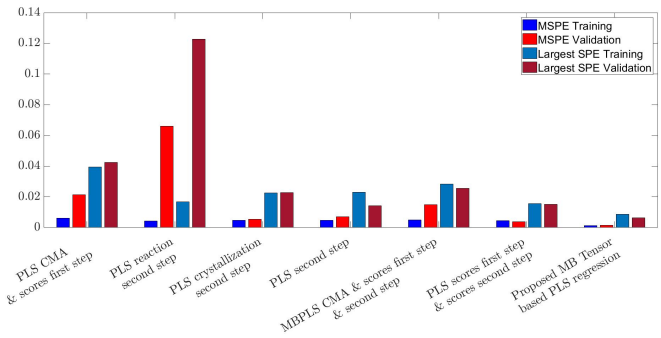

Fig. 4. Errors on the training and validation of alternative regression models compared to the application of the proposed Multi-block tensor regression.

of the first step and the CMAs of the inputs. Secondly, a tensor based PLS regression was trained based on the scores for the two synthesis steps but without the CMAs, and with no feature selection being applied. The results were compared based on the mean squared predictive error (MSPE) and the largest square predictive error both for the training and the validation data. These results are summarized in Fig. 4.

These results not only confirm the need for a multi-block regression to predict the final quality of the API but also prove that the proposed strategy outperforms other traditional methods. As it can be seen in Fig. 4 the first two PLS models, which are based only on the process conditions of the first synthesis step or the reaction on the second synthesis step have the worst performance compared to all the other models. From the large errors on the validation data, it is clear that these two models are unable to capture the correlations between inputs and output and therefore the low error on the training data is the result of over-fitting. The PLS model based on the crystallization during the second synthesis step produces significantly better results. A low MSPE with consistent results for the training and validation data demonstrates that in this case there is no over-fitting and also shows the strong correlation between this part of the process and the output. However, the results obtained for some of the multi-block models demonstrate that these results can still be improved if the correlations exiting with other variables in the input space are considered. The traditional MBPLS applied on the full (unfolded) data set does not result in a significant improvement compared to the PLS model on the crystallization of the second synthesis step. In fact, the errors in the validation data set are slightly higher. The last two results correspond to the regression models applying the proposed approach. First, an incomplete application that does not include the CMAs data and with no feature selection produces the second-best performance. This is surpassed only by the last model which is the complete application of the proposed approach to the full data set. This shows that the proposed method is more suitable to identify and capture the relevant correlations between the full input space and the output achieving high accuracy on the prediction.

The detailed results on the application of the Multi-block model to predict the quality attribute for the API are depicted in Figs. 5 and 6. The results for the training

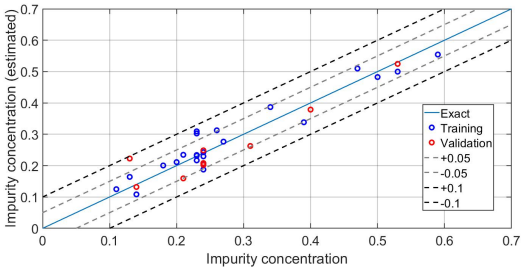

Fig. 5. Estimation of the impurity concentration for the training and validation batches

data and validation are plotted together. Fig. 5 depicts the relation between the actual quality attribute measured analytically on the API and the prediction made by the multi-block regression model. The impurity concentration in the API varies around the mean value of $\mu = 0.2772$ with $\sigma = 0.127$. This distribution is still considered normal variation because the impurity concentration is below the specification limit in every case. However, the model developed is highly relevant to gain an understanding of the phenomena that result in the variable presence of this impurity. This model can be then used to estimate the impurity concentration on new batches but most importantly to implement control actions that guarantee low levels of impurity reducing the risk of future production resulting on out of specifications product. Fig. 5 shows that estimations on training and validation have similar accuracy through the full range of variation of the predicted variable, with an overall correlation coefficient between the measurement and the estimation of $R^2 = 0.898$.

The box plots in fig. 6 display the absolute errors on the training and validation data. These results show that the errors on the training data are symmetrically distributed along the full range of the impurity concentration. This is evidence of the validity of the model for the estimation of this quality attribute of the API. The errors on the validation data set are in the same order of magnitude of the training errors. The partial bias observed towards positive errors in the validation data might be the result of the limited number of batches in the validation data set and therefore a larger data set must be evaluated to further verify the robustness of the model. The largest relative error obtained in the estimation is equivalent to $19,3\%$ of the range of variation of impurity encountered in the samples. Although the error is still large, this is the best accuracy archived from all the methods evaluated. Additionally, for the purpose of the model, this accuracy is sufficient since the prediction is not intended to replace the measurement of the impurity concentration but instead to generate an estimate on what can be the expected value given the process conditions and more importantly to understand the conditions that result on a higher presence of the impurity.

5.3 Root cause analysis

The multi-block tensor regression model was used to identify the process conditions or input material properties more correlated with the high values on the quality attribute for the API. This type of root cause analysis is intended to determine which conditions should be avoided

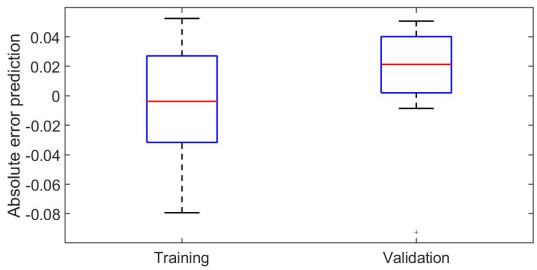

Fig. 6. Error distribution on the predictions for the training and validation batches.

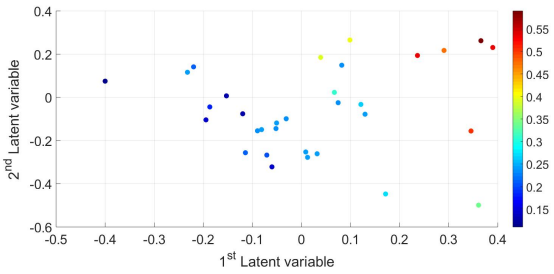

Fig. 7. Scatter plot for the first two latent variables of the multi-block regression model. Color scale given by the impurity concentration in every batch.

to reduce at most the risk of the product resulting in out of specifications. For the studied case, an out of specification occurs when the impurity concentration is higher than the quality specification for the API. The scatter plot depicted in Fig. 7 shows the scores for all batches in the data set. Only the first two, out of four, latent variables of the regression model are displayed. Every dot in the scatter plot represents one batch and the color is given by the corresponding impurity concentration on the product. This figure shows how the direction of major variation for the quality attribute is highly correlated with the change on the first two latent variables of the model. Thus, most of the batches with high impurity levels in the product are cluster at the top right corner of the latent space. On the contrary, those batches that resulted in the lowest levels of impurity are concentrated at the left side of the space. This separation indicates that the model has captured correlations that allow discriminating the product based on the quality attribute. In turn, this suggests that the model can help to elucidate causes for the differences observed in the product quality.

Relative contribution plots were used to identify the major contributions from the input space to high impurity levels. The relative contributions were computed for the difference between the mean of the batches that have a high impurity level and the mean for those batches around the centre of the latent space. Fig. 8 shows the relative contributions from each of the 18 selected features in the regression model. Large relative contributions can be interpreted as the major differences between the two means. The colors in this bar plot correspond with the different process steps or CMAs to which every feature belongs. Fig. 8 shows that the major contribution comes from the fifth latent variable of the factorized crystallization step in the second synthesis step. Other less significant relative contributions to a high

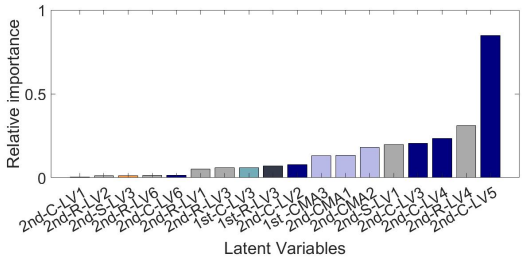

Fig. 8. Relative contributions in the regression model to a high impurity level.

impurity come from latent variables of the reaction in the second synthesis step and input material properties for the first synthesis step. The relative contributions from other selected features are not significant in the discrimination between high and low quality attribute for the API.

The process contributions to the feature identified as most relevant from Fig. 8 are independently computed from the Tucker3 factorization of the crystallization of the second synthesis step. Fig. 9 shows the contributions from some of the process variables measured during crystallization. The contributions to this latent variable show that the variance originates from a phenomenon that affects the temperature in the reactor and is perceived in every temperature sensor. The detailed profiles along time for these contributions were investigated to identify the range of time at which the variance originates. This allowed to identify a range of time around the middle of the crystallization operation as the source of variance, i.e., during anti-solvent addition,. Fig. 10 shows the trends for the vapor temperature during the crystallization. The red lines correspond to the batches that resulted on high impurity, while blue and black lines are all other batches. In this figure it can be observed that around the middle of the operation there is a difference in the trend of the vapor temperature between the two groups of batches. A sudden drop on the vapor temperature around the time point = 150 is observed only on those batches that resulted on a higher impurity level. This finding corresponds with observations made during the operation. The batches that experienced the drop in the vapor temperature had a faster crystallization once the anti-solvent addition started. This drop in temperature was also observed in other temperature sensors, but the largest variance is seen on the vapor temperature. Thus, the finding on the phenomena correlated with the high impurity, thanks to the model contributions, is in fact the evidence for the root cause that was a faster crystallization during the anti-solvent addition.

6. CONCLUSIONS

This contribution presents a novel strategy for multi-block tensor regression using hierarchical modelling. Batch process data is factorized via the simultaneous scaling and training of Tucker3 models to be the inputs for a PLS regression model. This strategy considers feature selection as a key step to identify and capture meaningful correlations between input and output spaces. The proposed strategy was applied to data from the large scale production of an API. Two synthesis steps of this production process were included in the modelling efforts. It was demonstrated

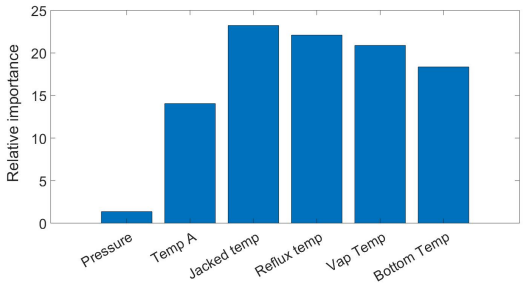

Fig. 9. Contributions to the fifth latent variable of the crystallization during the second synthesis step.

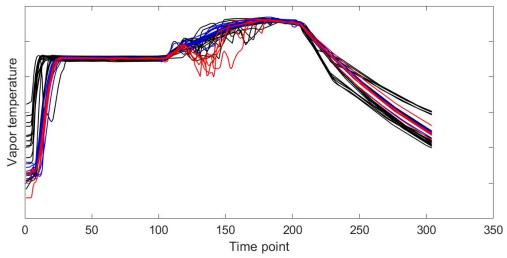

Fig. 10. Trends for the vapor temperature during the crystallization in the second synthesis step of the API.

that the proposed strategy performs better than other models based on single-unit modelling and other multi-block regression approaches. The novel strategy resulted in the lowest error for training and validation batches, with no sign of over-fitting, equally distributed error along with the full range of the output variable and equivalent errors for model training and validation. The model contributions helped in the investigation of the root cause for a higher impurity in the product. The combination of observations during operation and the contributions captured in the regression model resulted in the identification of the root cause. The process conditions during anti-solvent addition were adjusted to prevent the future occurrence of the uncontrolled fast crystallization. The future work aims at investigating how scalable is this approach in relevant industrial cases where more synthesis and sub-steps are involved.

ACKNOWLEDGEMENTS

This work was supported by Fonds Wetenschappelijk Onderzoek Vlaanderen[G.0863.18],the Belgian Science Policy Office(DYSCO)[IAPVII/19]and KULeuven PFV/10/002 Center-of-Excellence Optimizationin Engineering(OPTEC). CAM holds a VLAIO-Baekeland[HBC.2017.0239] grant.

REFERENCES

Biancolillo, A., Liland, K.H., Måge, I., Næs, T., and Bro, R. (2016). Variable selection inmulti-block regression. *Chemometrics and Intelligent Laboratory Systems*, 156, 89–101.

Biancolillo, A., Marini, F., and Roger, J.M. (2019). SO-CovSel: A novel method for variable selection in a multiblock framework. *Journal of Chemometrics*, 34, 1–21.

Ermi, B., Acar, E., and Cemgil, A.T. (2013). Link prediction in heterogeneous data via generalized coupled tensor factorization. *Data Mining and Knowledge Discovery*, 29, 203–236.

Fanaee, H. and Gama, J. (2016). Tensor-based anomaly detection: An interdisciplinary survey. *Knowledge-based systems*, 98, 130–147.

Gurden, S.P., Westerhuis, J.A., Bijlsma, S., and Smilde, A.K. (2001). Modelling of spectroscopic batch process data using grey models to incorporate external information. *Journal of Chemometrics*, 15, 101–121.

Hong, J.J., Zhang, J., and Morris, J. (2014). Progressive multi-block modelling for enhanced fault isolation in batch processes. *Journal of Process Control*, 24, 13–26.

Hu, K. and Yuan, J. (2009). Batch process monitoring with tensor factorization. *Journal of Process Control*, 19, 288–296.

Jiang, Q. and Yan, X. (2014). Plant-wide process monitoring based on mutual informationmultiblock principal component analysis. *ISA Transactions*, 53(5), 1516–1527.

Jiang, Q. and Yan, X. (2015). Nonlinear plant-wide process monitoring using MI-spectral clustering and Bayesian inference-based multiblock KPCA. *Journal of Process Control*, 32, 38–50.

Kourti, T., Nomikos, P., and MacGregor, J.F. (1995). Analysis, monitoring and fault diagnosis of batch processes using multiblock and multiway pls. *Journal of Process Control*, 5(4), 277–284.

Luo, L., Bao, S., Mao, J., and Tang, D. (2016). Quality prediction and quality-relevant monitoring with multilinear PLS for batch processes. *Chemometrics and Intelligent Laboratory Systems*, 150, 9–22.

MacGregor, J., Jaeckle, C., Kiparissides, C., and Koutoudi, M. (1994). Process monitoring and diagnosis by multiblock pls methods. *AIChE Journal*, 40(5), 826–838.

Måge, I., Smilde, A.K., and van der Kloet, F.M. (2018). Performance of methods that separate common and distinct variation in multiple data blocks. *Journal of Chemometrics*, 33, 1–3.

Munoz, C.A., Telen, D., Nimmegeers, P., and Van Impe, J. (2018). Feature extraction for batch process monitoring and fault detection via simultaneous data scaling and training of tensor based models. *IFAC-PapersOnLine*, 51(24), 433–440.

Nomikos, P. and MacGregor, J. (1994). Monitoring of batch processes using multi-way principal component analysis. *AIChE Journal*, 40, 1361–1375.

Sidiropoulos, N.D., De Lathauwer, L., Fu, X., Huang, K., Papalexakis, E.E., and Faloutsos, C. (2017). Tensor decomposition for signal processing and machine learning. *IEEE Transactions on signal processing*, 65, 3551–3582.

Skotare, T., Sjögren, R., Surowiec, I., Nilsson, D., and Trygg, J. (2018). Visualization of descriptive multiblock analysis. *Journal of Chemometrics*, 34, 1–11.

Tomba, E., Facco, P., Bezzo, F., and Barolo, M. (2013). Latent variable modeling to assist the implementation of Quality-by-Design paradigms in pharmaceutical development and manufacturing: A review. *International Journal of Pharmaceutics*, 457(1), 283–297.

Westerhuis, J.A. and Coenegracht, P.M. (1997). Multivariate modeling of the pharmaceutical two-step process of wet granulation and tableting with multiblock partial least squares. *Journal of Chemometrics*, 11(5), 379–392.